Zur Konstruktion fast-optimaler Algorithmen der Numerik

Braß, Helmut

Veröffentlicht in:
Abhandlungen der Braunschweigischen Wissenschaftlichen Gesellschaft Band 46, 1995, S.71-78

Verlag Erich Goltze KG, Göttingen

Zur Konstruktion fast-optimaler Algorithmen der Numerik

Von **Helmut Braß***, Braunschweig

(eingegangen am 14.12.1995)

1. Einleitung

Zur Behandlung der Standard-Probleme der Numerik, wie etwa Quadratur, sind im Lauf der Zeit viele Algorithmen vorgeschlagen worden. Will man die Entscheidung für einen speziellen Algorithmus nicht allein mit Tradition oder der Meinung von Autoritäten begründen, so hat man Qualitätsmaße zu definieren. Anhand eines solchen Maßes lassen sich Algorithmen vergleichen und der Begriff des "optimalen" Algorithmus kann präzisiert werden. Die Existenz optimaler Algorithmen ist für viele Situationen gesichert, jedoch begegnet ihre explizite Bestimmung erheblichen Schwierigkeiten (für eine hier behandelte Problemklasse siehe z.B. Köhler [4]). Aus diesen und anderen Gründen (z.B. Davis/Rabinowitz [3] p. 331, 332) werden optimale Algorithmen in der numerischen Praxis nur selten verwendet. Aber auch bei Anwendung nicht-optimaler Verfahren müssen Qualitätsmaße Leitlinie der Auswahl sein. Günstig wären Verfahren, die sowohl einfach zu konstruieren als auch nahezu optimal sind. Wir behandeln diese Aufgabe hier für eine spezielle, aber praktisch wichtige Klasse von Problemen.

Es bezeichne $C[0,1]$ den normierten Raum der auf $[0,1]$ definierten reellen stetigen Funktionen, versehen mit der sup-Norm. I sei ein auf $C[0,1]$ definiertes stetiges Funktional. Das zu behandelnde Problem besteht in der näherungsweisen Bestimmung von $I[f]$, wenn über f allein bekannt ist

(i) die Funktionswerte $f(\frac{\nu}{n}) \quad \nu = 0,1,\ldots,n$

(ii) $\|f^{(r)}\| \leq 1$

mit festen Zahlen n und r. (Die Ableitung $f^{(r)}$ ist im verallgemeinerten Sinne zu verstehen, d.h. $f^{(r-1)} \in C[0,1]$ und $|f^{(r-1)}(x) - f^{(r-1)}(y)| \leq |x-y|$). Die Festlegung der über f verwendeten Information (also hier (i) und (ii)) ist unumgänglich, wenn man verschiedene Algorithmen vergleichen will.

Ein Schätzwert für $I[f]$ wird bestimmt mit Hilfe eines Funktionals Q der Form

$$Q[f] = \sum_{\nu=0}^{n} a_\nu f\left(\frac{\nu}{n}\right)$$

* Prof. Dr. H. Braß · Institut für Angewandte Mathematik der TU Braunschweig
Pockelsstraße 14 · 38106 Braunschweig

mit passend gewählten Zahlen $a_0, \ldots, a_n$. Allgemeinere Ansätze für Q würden nicht mehr bringen, das zeigt der Satz von Smolyak (Traub/Woźniakowski [5] p. 54). Wir messen die Brauchbarkeit von Q durch

$$\rho(Q) := \sup\Big\{|I[f] - Q[f]| \quad : \quad \|f^{(r)}\| \leq 1\Big\}$$

Wir definieren weiter

$$\rho^{opt} := \inf\Big\{\rho(Q) \,:\, (a_0, \ldots, a_n) \in \mathbb{R}^{n+1}\Big\}$$

und haben als unser Qualitätsmaß

$$\mathrm{qual}(Q) := \frac{\rho(Q)}{\rho^{opt}}.$$

Vielfach werden Schätzformeln Q durch Anwendung von I auf einen Interpolationsausdruck $\mathrm{intpol}[f]$ konstruiert:

$$Q[f] = I \circ \mathrm{intpol}[f]$$

$$\mathrm{intpol}[f] := \sum_{\nu=0}^{n} f\left(\frac{\nu}{n}\right) l_\nu \quad , \quad l_\nu \in C[0,1],$$

$$l_\nu\left(\frac{\mu}{n}\right) = \delta_{\nu\mu}.$$

Für derartige Q läßt sich die Qualität leicht abschätzen, es gilt nämlich

Satz

$$\mathrm{qual}(I \circ \mathrm{intpol}) \quad \leq 1 + A$$

mit

$$A := \sup\Big\{\|\,\mathrm{intpol}[f]^{(r)}\| \,:\, \|f^{(r)}\| \leq 1\Big\}.$$

Beweis:

$$\begin{aligned}
\rho(I \circ \mathrm{intpol}) &= \sup\{|I[f - \mathrm{intpol}\, f]| \quad : \quad \|f^{(r)}\| \leq 1\} \\
&= (1+A)\sup\Big\{\Big|I\Big[\frac{f - \mathrm{intpol}[f]}{1+A}\Big]\Big| \,:\, \|f^{(r)}\| \leq 1\Big\} \\
&\leq (1+A)\sup\Big\{|I[g]| \,:\, \|g^{(r)}\| \leq 1 \text{ und } g(\tfrac{\nu}{n}) = 0, \nu = 0, \ldots, n\Big\} \\
&\leq (1+A)\rho^{opt}.
\end{aligned}$$

Es kommt hiernach darauf an, Interpolationsarten mit kleinen Werten von A aufzufinden. Polynominterpolation ist hierzu, jedenfalls für größere n – Werte, nicht brauchbar, was ja etwa durch die sehr schlechte Qualität des Newton-Cotes-Verfahrens (hierzu Braß [1]) belegt wird.

Eine Alternative bietet die Spline-Interpolation. Im Fall $r = 1$ ergibt lineare Spline-Interpolation (Spline-Knoten νn^{-1}) den bestmöglichen Wert $A = 1$, wie leicht zu sehen. Ziel dieser Arbeit ist der Nachweis, daß man für $r = 2$ mit quadratischer Spline-Interpolation $A < 1{,}5$ und für $r = 3$ mit kubischen Splines $A < 1{,}59$ erhält. Damit sind für $r = 1, 2, 3$ einfache Verfahren zur Konstruktion fast-optimaler Algorithmen gegeben.

Satz 1 $\mathrm{intpol}_{2,n}[f]$ bezeichne den f an den Stützstellen ν $(\nu = 0,1,\ldots,n)$ interpolierenden quadratischen Spline mit den Knoten $\xi_\nu = \nu + \frac{1}{2}$ $(\nu = 1,2,\ldots,n-2)$. Es gilt

$$\sup_n \sup\left\{\left\|\mathrm{intpol}_{2,n}[f]''\right\| : \|f''\| \le 1\right\} = \frac{3}{2}.$$

Satz 2 $\mathrm{intpol}_{3,n}[f]$ bezeichne den f an den Stützstellen νn^{-1} $(\nu = 0,1,\ldots,n)$ interpolierenden kubischen Spline mit den Knoten $\xi_\nu = \nu$ $(\nu = 2,3,\ldots,n-2)$. Es gilt

$$\sup_n \quad \sup\{\|\mathrm{intpol}_{3,n}[f]'''\| \quad : \quad \|f'''\| \le 1\} =$$
$$1 + \frac{(\sqrt{2}+1)(\sqrt{3}-1)}{3} \quad = \quad 1{,}589\ldots.$$

Es sei noch hinzugefügt, daß die angegebenen Suprema schon für kleine n nahezu erreicht werden, so erhält man für $n = 6$ im quadratischen Fall $1{,}492\ldots$ und im kubischen Fall $1{,}531\ldots$.

Es wäre von großem Interesse, diese Resultate auf höhere Ableitungen auszudehnen, jedoch stößt die hier benutzte Methode auf technische Schwierigkeiten. Beim analogen Fall periodischer Funktionen lassen sich diese jedoch überwinden, siehe dazu Braß [2].

2. Hilfssätze über eine Zahlenfolge

Wir stellen hier einige unten benötigte Resultate über eine rekursiv definierte Folge a_ν zusammen.

Es sei

$$a_0 = a_1 = 1$$
$$a_{\nu+1} + 2a\,a_\nu + a_{\nu-1} = 0 \quad \nu = 1,2,\ldots \tag{1}$$

mit einem $a > 1$. Mit bekannten Methoden erhält man den expliziten Ausdruck

$$a_\nu = \alpha\lambda_1^\nu + \beta\lambda_2^\nu$$

$$\lambda_1 = -a + \sqrt{a^2-1}, \qquad \lambda_2 = -a - \sqrt{a^2-1},$$

$$\alpha = \frac{1}{2}\left(1 + \sqrt{\frac{a+1}{a-1}}\right), \qquad \beta = \frac{1}{2}\left(1 - \sqrt{\frac{a+1}{a-1}}\right).$$

Auf Grund dieser Formeln lassen sich die folgenden Beziehungen unschwer verifizieren:

$$|a_{\nu-1}| \quad \le \quad |a_\nu| \qquad \nu = 1,2,\ldots \tag{2}$$

$$\operatorname{sgn} a_\nu = (-1)^{\nu-1} \qquad \nu = 1,2,\ldots \tag{3}$$

$$a_p a_{q+1} - a_{p-1} a_q = a_{p+q} - a_{p+q-1} \tag{4}$$

$$a_p^2 + 2a\, a_p\, a_{p+1} + a_{p+1}^2 = 2a + 2 \tag{5}$$

$$\sum_{\nu=1}^{\tau} |a_\nu| = \frac{|a_{\tau+1}| - |a_\tau| - 2}{2a - 2} \tag{6}$$

$$\sum_{\nu=2}^{\infty} \frac{1}{|a_\nu|} \leq \frac{1}{|a_2|(1 - a + \sqrt{a^2 - 1})} \tag{7}$$

$$|a_p\, a_q| \leq \frac{1}{2\sqrt{a^2 - 1}} |a_{p+q} - a_{p+q-1}| \tag{8}$$

$$\lim_{\substack{p\to\infty \\ q\to\infty}} \left| \frac{a_p\, a_q}{a_{p+q} - a_{p+q-1}} \right| = \frac{1}{2\sqrt{a^2 - 1}} \tag{9}$$

$$\left| \frac{a_{\nu+1}}{a_\nu} \right| > a + \sqrt{a^2 - 1} \qquad \nu = 1,2,\ldots \tag{10}$$

3. Beweis von Satz 2

Man sieht leicht, daß die hier interessierende Zahl A invariant gegenüber affinen Transformationen der Stützstellen ist. Wir können daher im folgenden kubische Spline-Interpolation mit Stützstellen $0,1,\ldots,n$ und Knoten $2,3,\ldots,n-2$ betrachten. Es bezeichne s den die Funktion $f \in C[0,n]$ interpolierenden Spline. Mit dem Ansatz

$$s(x)\big|_{[\nu,\nu+1]} = \frac{1}{6} t_\nu (x-\nu)^3 + t_{\nu,1}(x-\nu)^2 + t_{\nu,2}(x-\nu) + f(\nu)$$

$$\nu = 0,1,\ldots,n-1$$

ist also

$$A = \sup_\nu \sup\{|t_\nu| : \|f'''\| \leq 1\}. \tag{11}$$

Wir bezeichnen die dividierte Differenz der Funktion g zu den Stützstellen $\nu, \nu+1, \nu+2, \nu+3$ mit $\operatorname{dvd}(\nu,\ldots,\nu+3)[g]$. Die Peano-Kern-Darstellung

$$\operatorname{dvd}(\nu,\ldots,\nu+3)[g] = \int_\nu^{\nu+3} g'''(u) B(u-\nu)\, du \tag{12}$$

mit

$$B(u) := -\frac{1}{12}(-u)_+^2 + \frac{1}{4}(1-u)_+^2 - \frac{1}{4}(2-u)_+^2 + \frac{1}{12}(3-u)_+^2$$

ist wohlbekannt. Setzt man s in (12) ein, so folgt wegen $f(\nu) = s(\nu)$

$$\begin{aligned} &\operatorname{dvd}(\nu, \ldots, \nu+3)[f] \\ &= t_\nu \int_\nu^{\nu+1} B(u-\nu)du \;+\; t_{\nu+1} \int_{\nu+1}^{\nu+2} B(u-\nu)du \;+\; t_{\nu+2} \int_{\nu+2}^{\nu+3} B(u-\nu)du \\ &= \frac{1}{36}(t_\nu + 4t_{\nu+1} + t_{\nu+2}). \end{aligned}$$

Mit der Notation $\delta_\nu := 36 \operatorname{dvd}(\nu, \ldots, \nu+3)[f]$ haben wir also das Gleichungssystem

$$\begin{aligned} t_\nu + 4t_{\nu+1} + t_{\nu+2} &= \delta_\nu \qquad \nu = 0, 1, \ldots, n-3 \\ t_0 &= t_1 \\ t_{n-1} &= t_{n-2} \quad , \end{aligned} \tag{13}$$

wobei die beiden letzten Gleichungen besagen, daß bei 1 und $n-1$ keine Knoten vorliegen.

Das System (13) kann explizit gelöst werden. Zu diesem Zweck zieht man die Folge a_ν aus Abschnitt 2 mit dem Parameterwert $a = 2$ heran. Für diesen Spezialfall wollen wir d_ν statt a_ν schreiben.

Man beweist nun durch Einsetzen unter Heranziehung von (4)

$$t_\sigma = \frac{1}{d_{n-2} - d_{n-1}} \left(d_{n-1-\sigma} \sum_{\lambda=0}^{\sigma-1} d_{\lambda+1}\delta_\lambda + d_\sigma \sum_{\lambda=\sigma}^{n-3} d_{n-2-\lambda}\delta_\lambda \right) \tag{14}$$

$$\sigma = 1, 2, \ldots, n-2.$$

Mit Hilfe von (12) folgt

$$(d_{n-2} - d_{n-1})t_\sigma \;=\; \int_0^n f'''(u) K_\sigma(u) du$$

mit

$$K_\sigma(u) \;=\; 36\, d_{n-1-\sigma} \sum_{\lambda=0}^{\sigma-1} d_{\lambda+1} B(u-\lambda) \;+\; 36\, d_\sigma \sum_{\lambda=\sigma}^{n-3} d_{n-2-\lambda} B(u-\lambda).$$

Damit haben wir wegen (11)

$$\left| d_{n-2} - d_{n-1} \right| A \;=\; \sup \left\{ \int_0^n |K_\sigma(u)| du \;:\; 1 \le \sigma \le n-2 \right\}. \tag{15}$$

Auf Grund der Bedeutung von t_σ ist

$$K_\sigma(u) \;=\; K_{n-1-\sigma}(n-u)$$

leicht zu verifizieren, somit ist

$$\begin{aligned}(16)\qquad &\int_0^n |K_\sigma(u)|du = \int_0^\sigma \ldots + \int_\sigma^{\sigma+1} \ldots + \int_{\sigma+1}^n \ldots \\ &= \int_0^\sigma |K_\sigma(u)|du + \int_\sigma^{\sigma+1} |K_\sigma(u)|du + \int_0^{n-\sigma-1} |K_{n-\sigma-1}(u)|du.\end{aligned}$$

Wir setzen nun $1 < \sigma < n-2$ voraus.
K_σ ist ein quadratischer Spline, von dem man leicht feststellt, daß er in $[\kappa, \kappa+1]$ $(\kappa = 1,2,\ldots,\sigma-1,\sigma+1,\sigma+2,\ldots,n-2)$ einen Zeichenwechsel hat. Zusammen mit den doppelten Nullstellen bei 0 und n ergeben sich so $n+1$ Nullstellen, das ist die Maximalzahl für einen quadratischen Spline (ohne Null-Intervalle) mit $n-1$ Knoten, somit hat K_σ auf $[\sigma,\sigma+1]$ festes Vorzeichen. Mit einer kurzen Rechnung unter Beachtung von (1) und (4) erhält man

$$(17)\qquad \int_\sigma^{\sigma+1} |K_\sigma(u)|du = \left|\int_\sigma^{\sigma+1} K_\sigma(u)du\right| = \left|d_{n-2} - d_{n-1}\right|.$$

Weiter erhält man mit teils etwas mühsamer Rechnung (insbesondere ist (5) anzuwenden)

$$\int_0^1 |K_\sigma(u)|du = |d_{n-1-\sigma}|$$

$$\int_1^2 |K_\sigma(u)|du = \frac{95+64\sqrt{2}}{49}|d_{n-1-\sigma}|$$

$$\int_\kappa^{\kappa+1} |K_\sigma(u)|du = \frac{2}{3}|d_{n-1-\sigma}|\left(\left|2d_\kappa + d_{\kappa-1}\right|\left|1-d_\kappa^{-2}\right| + \sqrt{6}\left|d_\kappa + d_\kappa^{-1}\right|\left(1+d_\kappa^{-2}\right)^{\frac{1}{2}}\right)$$

$$\kappa = 2,3,\ldots,\sigma-1.$$

Man vergrößert nun unter Beachtung von (2), (3), (10) gemäß

$$\begin{aligned}&\left|2d_\kappa + d_{\kappa-1}\right|\left|1-d_\kappa^{-2}\right| + \sqrt{6}\left|d_\kappa + d_\kappa^{-1}\right|\left(1+d_\kappa^{-2}\right)^{\frac{1}{2}} \\ &\le \left|2d_\kappa + d_{\kappa-1} - d_\kappa^{-1}\left(2 + d_{\kappa-1}d_\kappa^{-1}\right)\right| + \sqrt{6}\left|d_\kappa + d_\kappa^{-1}\right|\left(1+\frac{1}{2}d_\kappa^{-2}\right) \\ &\le \left|2d_\kappa + d_{\kappa-1} - \sqrt{3}d_\kappa^{-1}\right| + \sqrt{6}\left|d_\kappa + \frac{3}{2}d_\kappa^{-1} + \frac{1}{2}d_\kappa^{-3}\right| \\ &\le (2+\sqrt{6})|d_\kappa| - |d_{\kappa-1}| + \left(\sqrt{6}\left(\frac{3}{2}+\frac{1}{50}\right) - \sqrt{3}\right)|d_\kappa^{-1}|\end{aligned}$$

und erhält mit Hilfe von (6) und (7)

$$\begin{aligned}\sum_{\kappa=2}^{\sigma-1}\int_\kappa^{\kappa+1} |K_\sigma(u)|du \le \frac{2}{3}|d_{n-1-\sigma}|\Bigg(&\left|\frac{\sqrt{6}+1}{2}d_\sigma + \frac{\sqrt{6}-1}{2}d_{\sigma-1}\right| \\ &- 3 - 2\sqrt{6} + \left(\sqrt{6}\frac{38}{25} - \sqrt{3}\right)\frac{\sqrt{3}+1}{10}\Bigg)\end{aligned}$$

und schließlich

$$\int_0^\sigma |K_\sigma(u)|du \quad \leq \quad \frac{2}{3}|d_{n-1-\sigma}|\left|\frac{\sqrt{6}+1}{2}d_\sigma + \frac{\sqrt{6}-1}{2}d_{\sigma-1}\right|.$$

Es folgt mit (1) und (4)

$$\begin{aligned}
&\int_0^\sigma |K_\sigma(u)|du + \int_0^{n-\sigma-1} |K_{n-\sigma-1}(u)|du \\
&\leq \frac{2}{3}\left|(\sqrt{6}+1)d_\sigma d_{n-1-\sigma} + \frac{\sqrt{6}-1}{2}d_{\sigma-1}d_{n-1-\sigma} + \frac{\sqrt{6}-1}{2}d_\sigma d_{n-\sigma-2}\right| \\
&= \frac{2}{3}\left|(3-\sqrt{6})d_\sigma d_{n-1-\sigma} + \frac{\sqrt{6}-1}{2}\Big(d_\sigma d_{n-\sigma-2} - d_{\sigma+1}d_{n-1-\sigma}\Big)\right| \\
&= \frac{2}{3}\left|(3-\sqrt{6})d_\sigma d_{n-1-\sigma} + \frac{\sqrt{6}-1}{2}\Big(d_{n-2} - d_{n-1}\Big)\right|.
\end{aligned}$$

Durch Anwendung von (8), (16) und (17) ergibt sich

$$\int_0^n |K_\sigma(u)|du \leq \left(1 + \frac{(\sqrt{3}-1)(\sqrt{2}+1)}{3}\right)|d_{n-1} - d_{n-2}| = |d_{n-1} - d_{n-2}| \cdot 1,58\ldots$$

Es bleiben noch die Fälle $\sigma = 1$ und $\sigma = n-2$ zu erörtern. Man erhält hier mit ähnlichen Rechnungen

$$\int_0^n |K_\sigma(u)|du \leq \frac{\sqrt{6}+2}{3}|d_{n-1} - d_{n-2}| < |d_{n-1} - d_{n-2}| \cdot 1,49$$

(15) zeigt nun

$$A \quad \leq \quad 1 + \frac{(\sqrt{3}-1)(\sqrt{2}+1)}{3}.$$

Diese Schranke kann nicht verbessert werden, wie man leicht erkennt, wenn man die im Beweis benutzten Majorisierungen nachprüft und insbesondere (9) beachtet.

4. Beweis von Satz 1

Der Beweis ähnelt dem vorhergehenden, wir fassen uns kurz. Man geht aus von dem Ansatz

$$s(x)\Big|_{[\nu-\frac{1}{2},\nu+\frac{1}{2}]} = \frac{1}{2}t_\nu(x-\nu)^2 + t_{\nu,1}(x-\nu) + f(\nu)$$
$$\nu = 0, 1, \ldots, n$$

und erhält an Stelle von (13)

$$\begin{aligned}
t_\nu + 6t_{\nu+1} + t_{\nu+2} &= \delta_\nu \qquad \nu = 0, 1, \ldots, n-2 \\
t_0 &= t_1 \\
t_{n-1} &= t_n \\
\delta_\nu &:= 16\ \mathrm{dvd}(\nu, \nu+1, \nu+2)[f].
\end{aligned}$$

Die Lösung läßt sich wiederum mit der Folge a_ν darstellen, diesmal zum Parameterwert $a = 3$. Wir schreiben e_ν für diese speziellen a_ν und haben

$$t_\sigma = \frac{1}{e_{n-1} - e_n}\left(e_{n-\sigma}\sum_{\lambda=0}^{\sigma-1} e_{\lambda+1}\delta_\lambda + e_\sigma \sum_{\lambda=\sigma}^{n-2} e_{n-1-\lambda}\delta_\lambda\right). \qquad \sigma = 1, 2, \ldots, n-1 .$$

Man erhält nach einiger Rechnung

$$\begin{aligned} &|e_n - e_{n-1}| \sup\left\{|t_\sigma| : \|f''\| \le 1\right\} \\ &= \frac{3}{2}|e_n - e_{n-1}| + |e_{n-\sigma}|\left(8\sum_{\lambda=1}^{\sigma-1}\frac{1}{|e_{\lambda+1} - e_\lambda|} - 2\right) \\ &\quad + |e_\sigma|\left(8\sum_{\lambda=1}^{n-\sigma-1}\frac{1}{|e_{\lambda+1} - e_\lambda|} - 2\right) \end{aligned}$$

woraus wegen

$$\sum_{\lambda=1}^{\infty}\frac{1}{|e_{\lambda+1} - e_\lambda|} \le \sum_{\lambda=2}^{\infty}\frac{1}{|e_{\lambda+1}|} < \frac{\sqrt{2}+1}{14} < \frac{1}{4}$$

alles folgt.

5. Dank

Ich danke Frau cand. math. A.-B. Eriksen für die sorgfältige Durchsicht des Manuskriptes.

6. Literatur

[1] Braß, H.: Quadraturverfahren. Vandenhoeck und Ruprecht 1977.
[2] Braß, H.: On the quality of algorithms based on spline interpolation. Zur Publikation eingereicht.
[3] Davis, P.J. und Rabinowitz, P.: Methods of numerical integration. Academic Press 1984.
[4] Köhler, P.: Optimale Quadraturformeln für Funktionen mit beschränkter zweiter Ableitung bei äquidistanten Stützstellen. In: Numerical Integration III (Eds.: Braß, H. und Hämmerlin, G.). Birkhäuser 1988.
[5] Traub, J.F. und Woźniakowski, H.: A general theory of optimal algorithms. Academic Press 1980.